IDIAP RESEARCH REPORT

AUTOMATIC ACCENTEDNESS EVALUATION OF NON-NATIVE SPEECH USING PHONETIC AND SUB-PHONETIC POSTERIOR PROBABILITIES

Ramya Rasipuram Milos Cernak
Alexandre Nanchen Mathew Magimai.-Doss

Idiap-RR-12-2015

JUNE 2015

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

AUTOMATIC ACCENTEDNESS EVALUATION OF NON-NATIVE SPEECH USING PHONETIC AND SUB-PHONETIC POSTERIOR PROBABILITIES

Ramya Rasipuram, Milos Cernak, Alexandre Nachen and Mathew Magimai-Doss

Idiap Research Institute, Martigny, Switzerland

{ramya.rasipuram, milos.cernak, alexandre.nanchen, mathew}@idiap.ch

ABSTRACT

Automatic evaluation of non-native speech accentedness has potential implications for not only language learning and accent identification systems but also for speaker and speech recognition systems. From the perspective of speech production, the two primary factors influencing the accentedness are the phonetic and prosodic structure. In this paper, we propose an approach for automatic accentedness evaluation based on comparison of instances of native and non-native speakers at the acoustic-phonetic level. Specifically, the proposed approach measures accentedness by comparing phone class conditional probability sequences corresponding to the instances of native and non-native speakers, respectively. We evaluate the proposed approach on the EMIME bilingual and EMIME Mandarin bilingual corpora, which contains English speech from native English speakers and various non-native English speakers, namely Finnish, German and Mandarin. We also investigate the influence of the granularity of the phonetic unit representation on the performance of the proposed accentedness measure. Our results indicate that the accentedness ratings by the proposed approach correlate consistently with the human ratings of accentedness. In addition, our studies show that the granularity of the phonetic unit representation that yields the best correlation with the human accentedness ratings varies with respect to the native language of the non-native speakers.

Index Terms: Automatic accent evaluation, non-native speech, phonetic representation, posterior features, KL-divergence, dynamic programming

1 Introduction

Non-native speakers of a language typically have accent because they tend to carry the phonetic and prosodic structure, and pronunciation rules from their mother tongue. Much of the research in accent evaluation relies on native speakers to listen to samples of accented speech and rate the accentedness. The goal of automatic accent evaluation is to identify the characteristics that contribute to a speakers accent automatically. Reliable and automatic evaluation of accentedness can provide many potential benefits to computer assisted language learning, second language acquisition research, accent identification, accent classification, speech and speaker recognition.

In the literature, automatic assessment of non-native speech has often focused on pronunciation error detection and assessment at the phoneme level. A variety of measures have been proposed to measure the pronunciation errors at phoneme level, such as log-likelihood based measures [1], log-likelihood ratio [2], goodness of pronunciation measure [3], log-posterior probability scores [1, 4], measures based on phonological features [5, 6]. Most of these measures are extracted from the output of an hidden Markov model (HMM) based speech recognizer. To improve the performance of mispronunciation detection, it is formulated as a 2-class classification task to determine if the pronunciation of a specific phoneme was 'correct' or 'wrong'. Many classifier-based approaches such as decision trees [7], linear discriminant analysis [8], logistic regression classifiers [9] are applied. However, the improvement in performance comes with a tradeoff that human annotations are required. On the other-hand, there are approaches to measure prosodic features of pronunciation [10, 11, 12, 13].

In this paper, we propose an approach to automatically evaluate the accentedness of non-native speakers (Section 2). The proposed approach compares the acoustic-phonetic content in native and non-native speakers speech in an instance-based framework. One of the main advantage of the proposed approach is its capability to go beyond instantaneous phoneme-level scoring, and provide utterance level and speaker level scoring of accentedness. Furthermore, the approach does not require any human labeled training data. The proposed approach is motivated from [14], where it was shown that the intelligibility of synthetic speech can be assessed objectively by comparing instances of synthetic speech and human reference speech.

We evaluate the proposed approach on the EMIME bilingual [15] and EMIME Mandarin bilingual [16] corpora, which contain English speech from native and non-native speakers (Section 3). The non-native speakers are from various native language backgrounds, namely Finnish, German and Mandarin. We study the impact of the granularity of the phonetic unit

This research was funded by the Commission for Technology and Innovation (CTI) on "Automatic scoring and adaptive pedagogy for oral language learning (ScoreL2)". The authors would like to thank Dr. Mirjam Wester for kindly sharing with us the human accentedness ratings of the EMIME bilingual and EMIME Mandarin bilingual corpora; and their colleague Raphael Ullmann for the fruitful discussions.

representation on the performance of automatic accentedness evaluation (Section 4). At the utterance level, the accentedness ratings by the proposed approach and the human ratings of accentedness correlate with a Pearson coefficient of 0.5 for Finnish and German speakers and 0.7 for Mandarin speakers.

2 Automatic Accentedness Evaluation

Recently [17, 18], it has been observed that the problem of matching a text hypothesis (typically represented as a sequence of lexical units) and an acoustic signal (typically a sequence of acoustic features) in an automatic speech recognition (ASR) system can be split into four sub-problems:

1. Definition of a latent symbol set.
2. Modeling the relationship between the acoustic feature observations and the latent symbols (acoustic model). Typically, Gaussian mixture models or artificial neural networks (ANNs) are used as acoustic models.
3. Modeling the relationship between the lexical units and the latent symbols (lexical model). This relationship can be deterministic (i.e., one-to-one) or probabilistic, and is typically learned or modeled using the acoustic model.
4. Matching of two sequences consisting of the evidence about the latent symbols from the acoustic model and the lexical model, respectively. This matching is typically performed using dynamic programming with local constraints and a local score that matches the acoustic and lexical models evidence at each time frame.

The present paper builds on this observation to show that objective evaluation of accentedness at acoustic-phonetic level and pronunciation level can be effectively formulated along the similar lines. That is, objective accentedness evaluation can be formulated as quantifying the mismatch between an acoustic signal representing the non-native speech and a text hypothesis representing the spoken message based on the knowledge of native speech. In the rest of the section, the four sub problems are elaborated in the context of objective accentedness evaluation.

Latent symbols: In an ASR system, latent symbols are typically based on phonemes, such as linguistic knowledge-driven context-independent phonemes or both linguistic knowledge and acoustic data-driven clustered context-dependent phonemes (obtained during decision tree-based state tying). In this paper, we show that the acoustic-phonetic differences between native and non-native speech can be better captured with clustered context-dependent phonemes derived using native speech data.

Acoustic model: Modeling the relationship between the acoustic feature observations and the latent symbols on native language speech data. In the paper, this relationship is modeled through artificial neural networks (ANNs).

Given an acoustic feature sequence $S = [\mathbf{s}_1, \cdots, \mathbf{s}_n, \cdots, \mathbf{s}_N]$ from a non-native speaker, the acoustic model estimates a latent symbol conditional probability vector sequence (posterior probability sequence) $Z = [\mathbf{z}_1, \cdots, \mathbf{z}_n, \cdots, \mathbf{z}_N]$, where N denotes the number of frames, and

$$\begin{aligned}\mathbf{z}_n &= [z_n^1, \ldots, z_n^k, \ldots, z_n^K]^{\mathrm{T}} \\ &= [P(c_1|\mathbf{s}_n), \ldots, P(c_k|\mathbf{s}_n), \ldots, P(c_K|\mathbf{s}_n)]^{\mathrm{T}}\end{aligned} \tag{1}$$

Here K denotes the number of latent symbols and $P(c_k|\mathbf{s}_n)$ denotes the posterior probability of latent symbol k given acoustic feature observation $\mathbf{s}_n$ at time frame n.

Lexical model: The lexical model can be linguistic knowledge-driven [19] or acoustic data-driven [20]. The first case represents the HMM-based approach and the second case represents the instance or template-based approach.

In the case of the linguistic knowledge-driven approach, a phonetic lexicon containing native pronunciation of words can be used and the text hypothesis, i.e., the text spoken by the non-native speaker is represented as a sequence of lexical units (context-independent phonemes or context-dependent phonemes). The relationship between lexical units and latent symbols can be captured through a decision tree (deterministic lexical model) or a probabilistic lexical modeling technique [19]. The probabilistic lexical model can be trained on native speech. In the case of deterministic lexical model, the relationship between each lexical unit in the sequence is a Kronecker delta distribution of cardinality K, in the case of probabilistic lexical model it is a categorical distribution of cardinality K. The lexical model eventually leads to a sequence of probability distributions $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_m, \cdots, \mathbf{y}_M]^{\mathrm{T}}$, where M is the lexical unit sequence or HMM state sequence length representing the text hypothesis.

In the case of the acoustic data-driven approach, the text hypothesis is represented through an instance or instances of speech signal(s) from a native speaker. In this case, given a acoustic feature observation sequence $X = [\mathbf{x}_1, \cdots, \mathbf{x}_m, \cdots, \mathbf{x}_M]^{\mathrm{T}}$ from a native speaker, the lexical model yields a sequence of latent symbol posterior probability vectors $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_m, \cdots, \mathbf{y}_M]^{\mathrm{T}}$ based on the acoustic model or an ANN. M in this case is the number of frames.

Matching of sequences Z and Y using dynamic programming: In dynamic programming, the local (temporal) constraints are dictated by the reference sequence Y. More precisely, when the lexical model is linguistic knowledge-driven, i.e., HMM-based, the accumulated score $A(m, n)$ is typically computed as

$$A(m, n) = S(\mathbf{y}_m, \mathbf{z}_n) + \min\left[A(m, n-1), A(m-1, n-1)\right]$$

where $S(\mathbf{y}_m, \mathbf{z}_n)$ is a local score that matches the acoustic model evidence $\mathbf{z}_n$ at time frame n with the lexical model evidence $\mathbf{y}_m$ at HMM state m. The local constraints here are the self transition of an HMM state ($A(m, n-1)$) and transition from the preceding state ($A(m-1, n-1)$). The above equation assumes uniform transition probabilities.

In the case of the instance-based approach, $A(m, n)$ can be computed using Itakura constraints, i.e.,

$$A(m,n) = S(\mathbf{y}_m, \mathbf{z}_n) + \min\big[A(m, n-1), A(m-1, n-1), A(m-2, n-1)\big]$$

The accumulated score $A(M, N)$ is normalized by the path length. The local match $S(\mathbf{y}_m, \mathbf{z}_n)$ can be based on a measure that compares two probability distributions. In this paper, we use symmetric Kullback-Leibler divergence, i.e.,

$$S(\mathbf{y}_m, \mathbf{z}_n) = \frac{1}{2} \cdot \sum_{k=1}^{K} \Big[z_n^k \log\Big(\frac{z_n^k}{y_m^k}\Big) + y_m^k \log\Big(\frac{y_m^k}{z_n^k}\Big)\Big] \tag{2}$$

The use of KL-divergence can be motivated from information-theoretic sense and hypothesis testing [21].

Given that the acoustic model (ANN) is trained on native speech and the lexical model is based on native speech data, it can be expected that $S(\mathbf{y}_m, \mathbf{z}_n)$ measures the instantaneous acoustic-phonetic mismatch, which when integrated over time through dynamic programming with local constraints yields pronunciation level mismatch between native and non-native speech. In other words, we hypothesize that the cumulative score $A(M, N)$ estimated in the proposed manner is a good measure of accentedness. We demonstrate that in the present paper using the instance-based lexical model approach.

In the literature [22], a combination of dynamic programming (using dynamic time warping) and classifier-based approaches was proposed for mispronunciation detection using spectral features and Gaussian posteriorgrams. However, measuring the acoustic-phonetic differences in spectral domain is difficult as it is more vunerable to undesirable variabilities such as speaker and environment. In this paper, the acoustic-phonetic differences between native and non-native speech are measured using latent symbol posterior probability sequences estimated through an ANN. ANNs are discriminative classifiers, and can provide invariance towards undesirable variabilities.

3 Experimental Setup

The experimental evaluations presented in this paper are conducted on the data from the EMIME bilingual [15] and EMIME Mandarin bilingual [16] corpora.

Speakers: The study consists of English speech from native and non-native speakers. We used the same three English utterances used in [15, 16, 23] spoken by all the speakers for which subjective accentedness ratings are available.

- *Native speakers:* 14 native English speakers consisting of seven male and seven female speakers from a variety of native accents namely, Scottish, Southern-English, American, New Zealand and Australian accents.
- *Non-native speakers:* 14 non-native English speakers from each native language group i.e., German, Finnish or Mandarin. Each language group consists of seven male and seven female speakers.

The native language of the three non-native language groups used in study is from different language families: English and German both belong to the Germanic branch of the Indo-European language family; Finnish belongs to the Finnic language family; Mandarin belongs to the Sino-Tibetan language family. English and German are closely related and share many features. For example, the phonemes of English and German are similar, as are stress and intonation patterns. Furthermore, the non-native English speakers learned English in a variety of places and from a variety of English-accented teachers.

Human accentedness ratings: Human accent ratings were collected at University of Edinburgh in sound isolated booths [15, 16]. Accent ratings were collected from 18 female and 10 male English monolingual listeners. The database also consists of accent ratings non-native listeners, however we have not used them in this study. The listeners were asked to score the degree of foreign accent for each utterance on a scale from 0 or "no foreign accent" to 6 "strong foreign accent". The listeners showed moderate to substantial inter-rater agreement [15, 23]. More analysis of accent ratings by native and non-native listeners is in [23].

MLPs: In this paper, we use multilayer perceptrons (MLP)s as phoneme posterior probability estimators. We use the Wall Street Journal (WSJ) corpus [24] as domain-independent data to train the MLPs. We used the SI-284 training data which contains approximately 80 hours of speech data (or 36,515 utterances from 284 speakers). Phoneme-based lexicon was obtained from the CMUDict pronunciation dictionary and consists of 40 context-independent phones including silence.

The input to all the MLPs is 39-dimensional perceptual linear prediction (PLP) features with a nine frame temporal context (i.e., four frame preceding and four frame following context). We use five-layer (one input, three hidden and one output layer) MLPs that are trained with the frame level cross entropy error criteria using the Quicknet software.Each hidden layer consisted of 2000 hidden units. The target labels for the MLPs were obtained from the HMM/GMM system.

We also investigate the influence of the granularity of the phonetic unit representation on the performance of the proposed accentedness measure. Therefore, various MLPs were trained which differed mainly in terms of the number of output units.

- *MLP-CI-40*: An MLP trained to classify 40 context-independent phones.

- *MLP-CD-N*: MLPs trained to classify N context-dependent phone states. The context-dependent phone states were obtained by decision tree-based state clustering of context-dependent phones in HMM/GMM framework. The different number of context-dependent phone states N ($N \in \{183, 419, 1013, 1915, 2832\}$) were obtained by varying the hyper parameters the state occupancy count and the log-likelihood threshold during decision-tree based state clustering.

Automatic accentedness evaluation: Accentedness scores are computed through the instance-based lexical model approach described in Section 2. In the literature, it has been argued that the performance of accent and pronunciation evaluation systems is higher at utterance and speaker levels than at word and phoneme levels [25]. Furthermore, speaker level accent assesment also demonstrates the extent of consensus among the ratings of various utterances. In this paper, we compute both utterance and speaker level accentedness scores as follows:

- *Utterance-level scores:* Accentedness scores are computed between an instance of non-native speech utterance and the seven native speech utterances from the same gender as the non-native speaker. The minimum of the resulting seven scores is considered as the accentedness score for that utterance and the particular speaker.
- *Speaker-level scores:* The utterance level accentedness scores for the three utterances of a speaker are averaged to obtain the speaker level accentedness score.

4 Results

The accentedness scores by the proposed approach are correlated (by computing the Pearson correlation coefficient) with the human ratings of accentedness at the utterance level and speaker level.

4.1 Utterance-level analysis

Table 1 presents the utterance level correlation between the accentedness scores by the proposed approach and the human accentedness ratings for Finnish, German and Mandarin non-native speaker utterances with increasing phonetic granularity.

The results in the first row of the table indicate that when context-independent phonemes are used, the proposed approach achieves higher correlation with respect to the human ratings for Mandarin speaker utterances followed by Finnish and German speaker utterances. For German speaker utterances, the context-independent phonemes yield very low correlation with respect to the human ratings. As discussed before, the reason for this could be that English and German belong to same language families and hence there are a number of aspects of German that help with the correct production of English. In the literature, it was observed that phoneme-level mispronunciation detection approaches when applied to non-native speech that is closer to native speech, result in poor correlation with the human ratings of accentedness [26, 27]. The results in first row of Table 1, inline with the literature, indicate that it is difficult to predict the accentedness of German speaker utterances than the accentedness of Mandarin speaker utterances.

It can be observed from Table 1 that the effect of granularity of the phonetic representation on the correlation depends on the native language of the non-native speakers. Particularly, for German speaker utterances, as the granularity of the phonetic representation increases, the correlation with human ratings also increases. However, for Finnish and Mandarin speaker utterances, the increase in the granularity of the phonetic representation did not result in significant increase in the correlation values.

Table 1. Correlation at utterance level between accentedness scores of the proposed approach and human ratings for Finnish, German and Mandarin non-native speakers.

# of latent symbols	Finnish	German	Mandarin
40	0.45	0.29	0.62
183	0.39	0.44	0.64
419	0.42	0.48	0.61
1013	0.46	0.53	0.64
1915	0.48	0.52	0.63
2832	**0.49**	**0.55**	**0.66**

4.2 Speaker-level analysis

Table 2 presents the speaker level correlation between the accentedness scores by the proposed approach and the human accentedness ratings for Finnish, German and Mandarin non-native speakers with increasing phonetic granularity. The results show that the proposed approach achieves higher correlation with human accent ratings at the speaker level than at the utterance level.

Similar to the utterance level correlation, results in the first row of the table indicate that when context-independent phonemes are used, the proposed approach achieves higher correlation with respect to the human ratings for Mandarin

Table 2. Correlation at speaker level between accentedness scores of the proposed approach and human ratings for Finnish, German and Mandarin non-native speakers.

# of latent symbols	Finnish	German	Mandarin
40	0.52	0.49	0.72
183	0.53	0.74	0.72
419	0.54	0.79	0.70
1013	0.54	0.79	0.72
1915	0.54	**0.80**	0.70
2832	**0.56**	**0.80**	**0.73**

speakers followed by Finnish and German speakers. For German speakers, as the granularity of the phonetic representation increases, the correlation with human ratings also increases, while for Finnish and Mandarin speakers, the increase in the granularity of the phonetic representation did not result in significant increase in the correlation values. The results indicate that the finer-granularity phonetic representation particularly improves the automatic accentedness evaluation performance of German speakers whose native language is closer to English.

The boxplots in Figure 1 show the z-scores using the human accentedness ratings and automatic accentedness ratings with the proposed approach (with 2832 clustered context-dependent phonemes) for female and male speakers, respectively. For both human ratings and automatic ratings, larger score indicates greater degree of accentedness. In the boxplots, FF represents Finnish female, GF represents German female, MF represents Mandarin female, FM represents Finnish male, GM represents German male and MM represents Mandarin male. Following observations can be made from Figure 1:

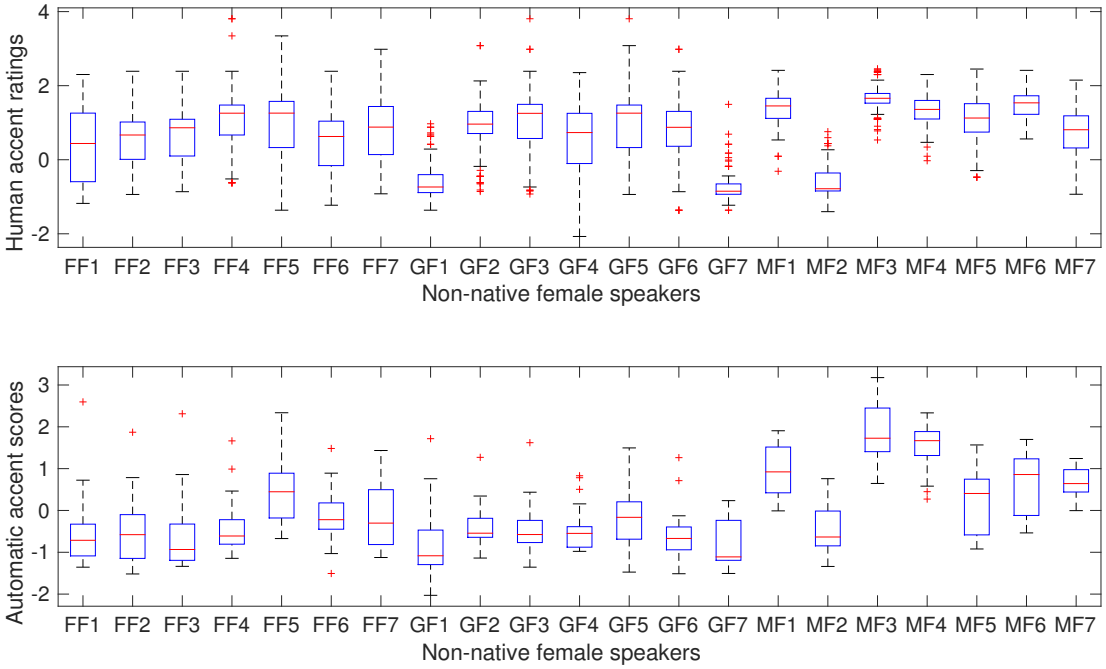
- FF5 is rated as most accented among Finnish female speakers by both human listeners and the proposed approach. FF1 is rated as least accented among Finnish female speakers by human listeners, however, the proposed approach identifies FF3 and FF1 as least accented.
- GF5 is rated as most accented among German female speakers by both human listeners and the proposed approach. GF1 and GF7 are rated as least accented among German female speakers by human listeners and the proposed approach.
- MF3 and MF6 were rated as highly accented among Mandarin female speakers by human listeners. The proposed approach identified only MF3 as highly accented among Mandarin female speakers. MF2 is rated as least accented among Mandarin female speakers by human listeners and the proposed approach.
- FM5 is rated as most accented by both human listeners and the proposed approach among Finnish male speakers.
- GM4 is rated as most accented among German male speakers by both human listeners and the proposed approach. GM1 is rated as least accented among German male speakers by both human listeners and the proposed approach.
- All Mandarin male speakers were given high accentedness ratings by the human listeners with a very smaller degree of variation [16]. The accentedness scores by the proposed approach also indicate relatively higher accentedness scores for all Mandarin male speakers compared to Finnish or German speakers.
- The least accented speakers among Finnish male speakers and Mandarin male speakers by the human listeners and proposed approach do not coincide.

Overall, the figures indicate that the speaker level trends in the accentedness ratings by humans and proposed approach are correlated.

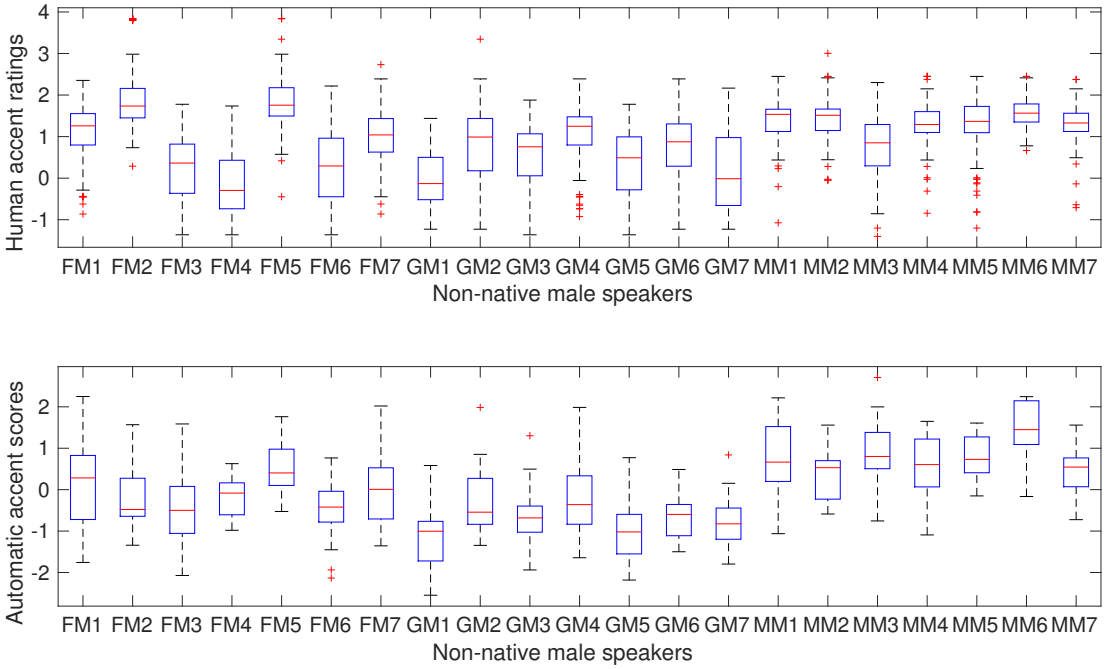
5 Conclusions

We proposed a novel approach for automatic accentedness evaluation of non-native speech based on comparison of acoustic-phonetic information obtained through an instance of non-native speech and the lexical and pronunciation information obtained through instances of native speech. We investigated the impact of the granularity of the phonetic unit representation on the performance of the proposed accentedness measure. Our investigations showed that the accentedness scores by the proposed approach correlate well with the human ratings of accentedness at both utterance level and speaker level. Furthermore, the granularity of the phonetic unit representation that results in optimal correlation with human accentedness ratings depends on the native language of the non-native speakers.

In this paper, the lexical and pronunciation structure were imposed through instances of native speech. As discussed in Section 2, this structure can also be imposed through a probabilistic lexical model trained on a native speech database through approaches such as Kullback-Leiber divergence based HMM. This provides two primary advantages. Firstly, it avoids the need for a reference native speech utterance. Secondly, it allows to localize individual phoneme or word level pronunciation errors and to provide a detailed error feedback to a non-native speaker or second language learner. Finally, a complete accent evaluation system should include assessment of prosodic characteristics of non-native accent. Our future work will focus on extending the approach to use linguistic knowledge-driven lexical model and to include prosodic features of accent.



(a) Female speakers



(b) Male speakers

Fig. 1. Accentedness scores based on human accent ratings and the proposed approach

6 References

[1] Y. Kim, H. Franco, and L. Neumeyer, "Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction," in *Proc. of EUROSPEECH*, 1997, pp. 64 564–64 568.

[2] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic Detection Of Phone-Level Mispronunciation For Language Learning," in *Proc. of EUROSPEECH*, 1999, pp. 851–854.

[3] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.

[4] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Proc. of Interspeech*, 2013, pp. 1886–1890.

[5] T. Stanley, K. Hacioglu, and B. Pellom., "Statistical machine translation framework for modeling phonological errors in computer assisted pronunciation training system," in *Proceedings of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2011.

[6] A. Sangwan and J. H. Hansen, "Automatic analysis of Mandarin accented English using phonological features," *Speech Communication*, vol. 54, no. 1, pp. 40–54, 2012.

[7] A. Ito, Y.-L. Lim, M. Suzuki, and S. Makino, "Pronunciation error detection for computer-assisted language learning system based on error rule clustering using a decision tree," *Acoustical Science and Technology*, vol. 28, no. 2, pp. 131–133, 2007.

[8] H. Strik, K. Truong, F. de Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[9] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[10] G.-A. Levow, "Investigating Pitch Accent Recognition in Non-Native Speech," in *Proceedings of ACL*, 2009.

[11] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners," in *Proceedings of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2009.

[12] A. Rilliard, A. Allauzen, and P. B. de Mareil, "Using Dynamic Time Warping to Compute Prosodic Similarity Measures," in *Proc. of Interspeech*, 2011, pp. 2021–2024.

[13] F. Hönig, A. Batliner, and E. Nöth, "Automatic Assessment of Non-Native Prosody - Annotation, Modelling and Evaluation," in *Proceedings of ISADEPT*, 2012.

[14] R. Ullmann, M. Magimai.-Doss, and H. Bourlard, "Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences," in *Proc. of ICASSP*, 2015.

[15] M. Wester, "The EMIME Bilingual Database," The University of Edinburgh, Tech. Rep. EDI-INF-RR-1388, 2010.

[16] M. Wester and H. Liang, "The EMIME Mandarin Bilingual Database," The University of Edinburgh, Tech. Rep. EDI-INF-RR-1396, 2011.

[17] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "On Modeling Context-Dependent Clustered States: Comparing HMM/GMM, Hybrid HMM/ANN and KL-HMM Approaches," in *Proc. of ICASSP*, 2014.

[18] M. Razavi and M. Magimai.-Doss, "On Recognition of Non-Native Speech Using Probabilistic Lexical Model," in *Proc. of Interspeech*, 2014.

[19] R. Rasipuram and M. Magimai.-Doss, "Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model," *Speech Communication*, vol. 68, pp. 23–40, 2015.

[20] S. Soldo, M. Magimai.-Doss, J. P. Pinto, and H. Bourlard, "Posterior Features for Template-based ASR," in *Proc. of ICASSP*, 2011.

[21] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. on Information Theory*, vol. IT-20, no. 4, 1974.

[22] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, Dec 2012, pp. 382–387.

[23] M. Wester and C. Mayo, "Accent rating by native and non-native listeners," in *Proc. of ICASSP*, 2014, pp. 7699–7703.

[24] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large Vocabulary Continuous Speech Recognition using HTK," in *Proc. of ICASSP*, vol. 2, 1994, pp. 125–128.

[25] S. M. Witt, "Automatic Error Detection in Pronunciation Training: Where we are and where we need to go," in *International Symposium on automatic detection on errors in pronunciation training*, 2012.

[26] P. Müller, F. D. Wet, C. V. D. Walt, and T. Niesler, "Automatically assessing the oral proficiency of proficient L2 speakers," in *Proceedings of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2009.

[27] K. Yan and S. Gong, "Pronunciation Proficiency Evaluation based on Discriminatively Refined Acoustic Models," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 3, no. 2, pp. 17–23, 2011.